

# Data 583 Life Expectancy (WHO)

Justin Chan, Kenny Tong, Viji Rajagopalan

22 Mar, 2023

## Data 583 Life Expectancy - Final Report (Life Expectancy Data)

### Part 1. Introduction and Hypotheses

Life Expectancy has always been an area of interest for humanity. The dataset contains the Life Expectancy records of 193 countries between 2000-2015, together with different predictive factors. Broadly speaking, predicting variables are categorized into 4 major areas : Immunization, Mortality, Economical, and Social, containing a total of 21 individual variables.

The primary purpose of this report is to compare and evaluate different predictive models in order to identify the most appropriate model for the dataset. In particular, we will evaluate the applicability of the core assumptions of the selected models by methods such as Normality Test, Multicollinearity Assessment, etc. This could validate or decline the adoption of certain models because the model assumptions are simply not satisfied. We are also going to verify whether the 4 predicting areas have equal significance on Life Expectancy, and whether there are adequate support evidence suggesting a strong correlation with the response variables. Finally, we will also perform and compare fitting result of selected models, in particular whether parametric models would be more suitable than non-parametric models for this dataset.

### Part 2. Dataset overview

#### Variables Types

Variable	Unit of Measurement/Data Category	Continuous vs Discrete
Life Expectancy	Years Old (Age)	Continuous
Country	Nominal Data	Discrete
Year	Ordinal Data	Discrete
Status	Nominal Data	Discrete
Adult Mortality	Count Data	Discrete
Infant deaths	Count Data	Discrete
Under-five deaths	Count Data	Discrete
Hepatitis B	Percentage	Continuous
Measles	Count Data	Discrete
Polio	Percentage	Continuous
Diphtheria	Percentage	Continuous
Total expenditure	Percentage	Continuous
Percentage expenditure	Percentage	Continuous
GDP	Currency (USD)	Continuous
Population	Count	Discrete
Income composition of resources	Percentage	Continuous
Schooling	Mean (Years)	Continuous

Variable	Unit of Measurement/Data Category	Continuous vs Discrete
Alcohol	Litres	Continuous
HIV/AIDS	Percentage	Continuous
BMI	Average BMI	Continuous
Thinness 1-19 years	Percentage	Continuous
Thinness 5-9 years	Percentage	Continuous

## Variables Summary and Categories

Life Expectancy is the response variable in this dataset. This represents the mean of the life expectancy (in age) in a specific country in a given year. For the data types of the predicting variables, most are percentage and count data across four major areas. The first area is Immunization Data such as Hepatitis B and Polio (immunization coverage). The second area is Mortality Data such as Adult Mortality/infant deaths (No. of deaths of Adult/infant per 1000 persons). The third area is Economical Data such as GDP/Income composition/Percentage expenditure. The fourth area is Social Data such as Schooling and Population.

To clean up and wrangle data for the subsequent analysis, NA data has been assessed. A total of 2563 NA values are found in the dataset, spreading across a few columns. These NA values are generally imputed by the respective column mean.

## Checking for Multicollinearity

As Multicollinearity can potentially affect the accuracy of regression model and we have 22 variables, a correlation study is undertaken to understand and assess the situation. A correlation plot has identified a number of correlation problems. It is found that infant deaths and under.five.deaths are nearly 100% correlated. The relation between the deaths rates of the two close age groups is easily interpretable. In addition, there are three heavily correlated pairs which is defined by the  $\text{abs}(\text{correlation coefficient}) > 0.7$  between the variables. They include (a) (immunization rate of) 'Polio'-vs-'Diphtheria', (b) 'income composition of resources'-vs-'Schooling', and (c) between the two thinness measures for the age groups 5-9 vs 10-19. Pairs (a) and (c) are justifiable while the relation for (b) demonstrate a relatively subtle relation. Other than that, the degree of multicollinearity is acceptable and not too worrying.



## Variables Selection

Due to the above correlations between certain variables, in addition to a large number of variables in our model, a variable selection is conducted to remove some correlated variables and attain a simpler model via variable selection. A BIC backward step model selection method has been applied to the dataset, with the following summary :

Models	No. of Variables	AIC Score	Adj R-squared Score
Original Model	20	7642.14	0.8299
Reduced Model	12	7604.24	0.8296

The reduced model now contains the following variables : Adult.Mortality + infant.deaths + Hepatitis.B + BMI + under.five.deaths + Polio + Diphtheria + HIV.AIDS + GDP + thinness..1.19.years + Income.composition.of.resources + Schooling

The number of independent variables are now effectively reduced to 12, together with a lower AIC score of 7604.34. Meanwhile, the adjusted R-squared score is well kept at nearly the same level as in the original model. We are satisfied with the performance of this reduced model. This reduced model will therefore be adopted as the basis for further analysis in this report.

## Part 3. Regression Analysis

PLEASE FEEL FREE TO MODIFY/RESTRUCTURE THE FOLLOWING SECTIONS ACCORDING TO THE FINAL REPORT

### Questions and Next Steps

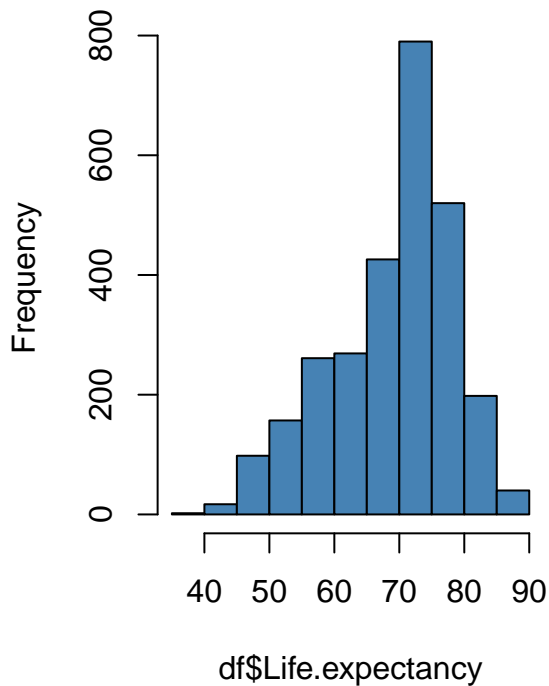
1. Does the variables selected using BIC and linear model able to explain Life Expectancy adequately? A hypotheses test is required for this.
2. Is the response variable normally distributed? Shapiro-Wilk test for normality will need to be conducted for this.
3. Is Multi linear regression the best model or go with other non parametric models? From initial feedback, there are hypotheses tests available to validate this.

Need to explore further on these questions from proposal stage and conclude.

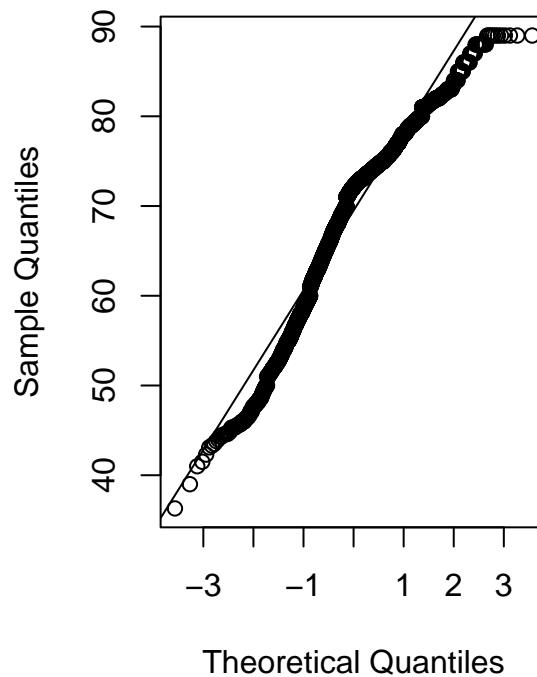
4. Understanding impact of individually controlled factors - The dataset has all predicting variables divided into 4 groups: Immunization related factors, Mortality factors, Economical factors and Social factors. Some of these factors are controllable by individuals like immunization, alcohol etc. Some of these factors are noncontrollable and macro elements like GDP. If an individual within a country want to improve life expectancy, how much is controllable/can be influenced personally? What proportion of variation in life expectancy can be explained by these variables? For example, What is the effect of “Alcohol/BMI” on the life expectancy?
5. Understanding impact of Government/Public controlled factors - From Government perspective, how are the preventive measures influencing life expectancy? What proportion of variation in life expectancy can be explained by these variables? For example, Does Higher health expenditure (column H) on Health improve life expectancy?
6. normal distribution test for our y variable

```
#Histogram & QQPlot
par(mfrow=c(1,2))
hist(df$Life.expectancy, col='steelblue', main='Life.expectancy_Histogram')
#not really a good "bell-shape"
qqnorm(df$Life.expectancy, main='Life.expectancy_QQplot')
#most of the data is not fall along a straight diagonal line
qqline(df$Life.expectancy)
```

Life.expectancy\_Histogram



Life.expectancy\_QQplot



*#Both are indicating that our predict variable Y "df\$Life.expectancy" is not normally distributed*

*#Shapiro-Wilk Test*

```
shapiro.test(df$Life.expectancy)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$Life.expectancy
## W = 0.95676, p-value < 2.2e-16
```

*#Finding: Since df\$Life.expectancy p-value is less than .05, indicate that our y variable is not normally*

*#Kolmogorov-Smirnov Test*

```
ks.test(df$Life.expectancy, 'pnorm')
```

```
## Warning in ks.test.default(df$Life.expectancy, "pnorm"): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  df$Life.expectancy
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

*#Finding: Since df\$Life.expectancy p-value is less than .05, indicate that our y variable is not normally*

LM with matching dependent variable with npreg

```
model_lm <- lm(Life.expectancy~Adult.Mortality + infant.deaths + Hepatitis.B + BMI + under.five.deaths + P
summary(model_lm)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +
##     Hepatitis.B + BMI + under.five.deaths + Polio + Diphtheria +
##     HIV.AIDS + GDP + thinness..1.19.years + Income.composition.of.resources +
##     Schooling, data = df, x = TRUE, y = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.0326  -2.1757  -0.1334   2.1880  18.8226
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.387e+01  5.251e-01 102.580 < 2e-16 ***
## Adult.Mortality   -1.709e-02  7.895e-04 -21.651 < 2e-16 ***
## infant.deaths      8.914e-02  7.908e-03  11.271 < 2e-16 ***
## Hepatitis.B       -1.382e-02  3.843e-03  -3.598 0.000327 ***
## BMI               3.848e-02  4.842e-03   7.947 2.76e-15 ***
## under.five.deaths -6.677e-02  5.817e-03 -11.477 < 2e-16 ***
## Polio             2.611e-02  4.501e-03   5.801 7.33e-09 ***
## Diphtheria        3.248e-02  4.764e-03   6.819 1.12e-11 ***
## HIV.AIDS          -4.808e-01  1.692e-02 -28.410 < 2e-16 ***
## GDP               6.234e-05  6.201e-06  10.054 < 2e-16 ***
## thinness..1.19.years -8.799e-02  2.241e-02  -3.927 8.83e-05 ***
## Income.composition.of.resources 6.728e+00  6.071e-01  11.082 < 2e-16 ***
## Schooling         8.110e-01  3.900e-02  20.793 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.884 on 2765 degrees of freedom
## Multiple R-squared:  0.8284, Adjusted R-squared:  0.8277
## F-statistic: 1113 on 12 and 2765 DF, p-value: < 2.2e-16
```

2. npreg on our dataset?

Note: When have time need to rerun with “x=True, y=True”, “VIF remove”, “add status.val”, “train split”

```
library(np)
```

```
## Nonparametric Kernel Methods for Mixed Datatypes (version 0.60-16)
## [vignette("np_faq",package="np") provides answers to frequently asked questions]
## [vignette("np",package="np") an overview]
## [vignette("entropy_np",package="np") an overview of entropy-based methods]
```

```
# n <- names(df)
# f <- as.formula(paste("df$Life.expectancy ~", paste(n[!n %in% "Life.expectancy"], collapse = " + ")))
```

```

#
# model_np <- npregbw(Life.expectancy ~ Adult.Mortality + infant.deaths + Hepatitis.B + BMI + under.five.d
# model_np <- npreg(bws = model_np)
# summary(model_np)
model_np <- readRDS("model_np.rds") #PreTrained Model
summary(model_np)

##
## Regression Data: 2778 training points, in 12 variable(s)
##      Adult.Mortality infant.deaths Hepatitis.B      BMI
## Bandwidth(s):      389457535      6733757    225092161 79216285
##      under.five.deaths  Polio Diphtheria HIV.AIDS      GDP
## Bandwidth(s):      95308072 5825954    19248839 1.393258 167351562078
##      thinness..1.19.years Income.composition.of.resources Schooling
## Bandwidth(s):      37667667      1071202 15165344
##
## Kernel Regression Estimator: Local-Linear
## Bandwidth Type: Fixed
## Residual standard error: 3.345092
## R-squared: 0.8722143
##
## Continuous Kernel Type: Second-Order Gaussian
## No. Continuous Explanatory Vars.: 12

# objects()
# find("model_np")
#
# saveRDS(model_np, "model_np.rds")

#npsigtest_npreg <- npsigtest(model_np)    #10 Hrs to run...

```

3. LASSO and Neuralnet Two different supervised algorithms tried on the dataset. They do not have the constraint of a normal distribution for response variable.

First did a train and test split so we can measure the MSE and compare how each of the models are performing interms of minimizing MSE.

```

library(glmnet)

## Warning: package 'glmnet' was built under R version 4.2.2

## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 4.2.2

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##      expand, pack, unpack

```

```
> npsigtest(model_np)

Kernel Regression Significance Test
Type I Test with IID Bootstrap (399 replications, Pivot = TRUE, joint = FALSE)
Explanatory variables tested for significance:
Adult.Mortality (1), infant.deaths (2), Hepatitis.B (3), BMI (4), under.five.deaths (5), Polio (6), Diphtheria (7), HIV.AIDS (8),
GDP (9), thinness..1.19.years (10), Income.composition.of.resources (11), Schooling (12)

Bandwidth(s): Adult.Mortality infant.deaths
               389457535      6733757
Bandwidth(s): Hepatitis.B      BMI under.five.deaths
               225092161 79216285      95308072
Bandwidth(s): Polio Diphtheria HIV.AIDS
               5825954 19248839 1.393258
Bandwidth(s): GDP thinness..1.19.years
               167351562078      37667667
Bandwidth(s): Income.composition.of.resources
               1071202
Bandwidth(s): Schooling
               15165344

Individual Significance Tests
P Value:
Adult.Mortality      < 2e-16 ***
infant.deaths        < 2e-16 ***
Hepatitis.B          0.047619 *
BMI                  < 2e-16 ***
under.five.deaths    < 2e-16 ***
Polio                < 2e-16 ***
Diphtheria           < 2e-16 ***
HIV.AIDS             < 2e-16 ***
GDP                  < 2e-16 ***
thinness..1.19.years < 2e-16 ***
Income.composition.of.resources < 2e-16 ***
Schooling            < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1: npsigtest\_npreg result

```
## Loaded glmnet 4.1-6
```

```
#70:30 split for train and test
```

```
df1<-df[,c('Life.expectancy','Adult.Mortality','infant.deaths','under.five.deaths',
           'Hepatitis.B','BMI','Polio','Diphtheria',
           'HIV.AIDS','thinness..1.19.years','Income.composition.of.resources','Schooling','GDP','Status.val')]

ind <- sample(1:nrow(df1), 2000)
traino <- df1[ind,]
testo <- df1[-ind,]
```

Linear model

```
lmmodtr <- lm(traino[,1]~.,data=traino[, -1],x=TRUE, y=TRUE)
summary(lmmodtr)
```

```
##
## Call:
## lm(formula = traino[, 1] ~ ., data = traino[, -1], x = TRUE,
##     y = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2091  -2.2438  -0.0971   2.3646  18.5510
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.437e+01  6.281e-01  86.567 < 2e-16 ***
```



```
## Adult.Mortality      -1.651e-02  9.268e-04 -17.818 < 2e-16 ***
## infant.deaths        8.874e-02  8.880e-03   9.993 < 2e-16 ***
## under.five.deaths    -6.664e-02  6.525e-03 -10.214 < 2e-16 ***
## Hepatitis.B          -1.357e-02  4.593e-03  -2.954  0.00317 **
## BMI                  4.064e-02  5.801e-03   7.007 3.33e-12 ***
## Polio                2.660e-02  5.138e-03   5.178 2.47e-07 ***
## Diphtheria           2.746e-02  5.518e-03   4.976 7.04e-07 ***
## HIV.AIDS             -4.864e-01  1.929e-02 -25.212 < 2e-16 ***
## thinness..1.19.years -5.912e-02  2.699e-02  -2.190  0.02861 *
## Income.composition.of.resources 6.370e+00 6.940e-01  9.179 < 2e-16 ***
## Schooling            7.796e-01  4.636e-02  16.819 < 2e-16 ***
## GDP                  5.114e-05  7.858e-06   6.508 9.62e-11 ***
## Status.val           1.484e+00  2.971e-01   4.995 6.40e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.93 on 1986 degrees of freedom
## Multiple R-squared:  0.8279, Adjusted R-squared:  0.8267
## F-statistic: 734.7 on 13 and 1986 DF,  p-value: < 2.2e-16
```

Another test to see if the above parametric model specification is correct.

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.2.2

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.2.2

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##      as.Date, as.Date.numeric
```

```
resettest(lmmodtr)
```

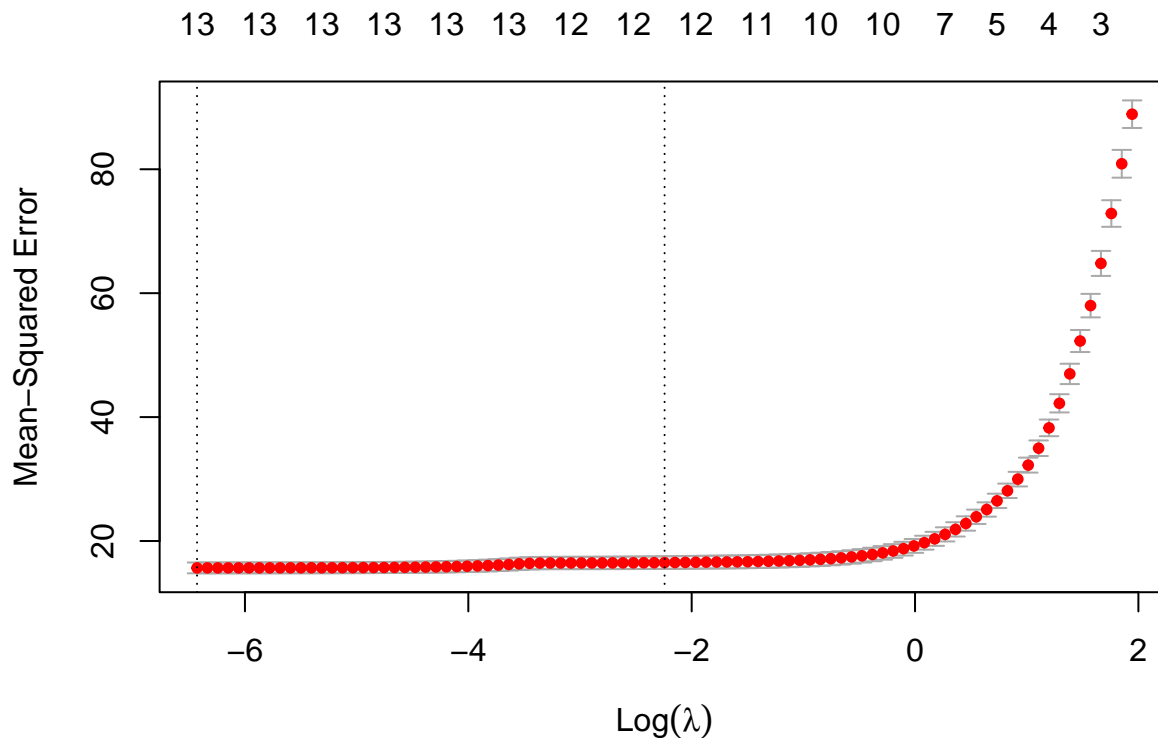
```
##
## RESET test
##
## data:  lmmodtr
## RESET = 87.863, df1 = 2, df2 = 1984, p-value < 2.2e-16
```

```
#LASSO
library(glmnet)
y <- traino$Life.expectancy
x <- data.matrix(traino[,-1])
#k-fold cross-validation to find optimal lambda value\
#cv default is 10 fold
cv_model <- cv.glmnet(x, y, alpha = 1)

#optimal lambda value that minimizes test MSE
best_lambda <- cv_model$lambda.min
best_lambda
```

```
## [1] 0.001613773
```

```
plot(cv_model)
```



```
#coefficients of best model
```

```
lasmod <- glmnet(as.matrix(traino[, -1]), traino$Life.expectancy, alpha = 1, lambda = best_lambda)
coef(lasmod)
```

```
## 14 x 1 sparse Matrix of class "dgCMatrix"
##                                     s0
## (Intercept)                       5.426894e+01
## Adult.Mortality                   -1.655666e-02
## infant.deaths                     8.070605e-02
## under.five.deaths                 -6.075701e-02
## Hepatitis.B                      -1.363371e-02
## BMI                              4.068818e-02
## Polio                            2.673394e-02
## Diphtheria                       2.797718e-02
## HIV.AIDS                        -4.872884e-01
## thinness..1.19.years             -5.665547e-02
## Income.composition.of.resources  6.432284e+00
## Schooling                        7.807371e-01
## GDP                             5.080599e-05
## Status.val                       1.463666e+00
```

```
#linear model
```

```
mselm_tel <- mean((testo[, 1] - predict(lmmodtr, newdata=testo))^2)
```

```
#lasso
mselas_tel <- mean((testo[,1]-predict(lasmod, newx=as.matrix(testo[,,-1])))^2)
print(mselm_tel)
```

```
## [1] 13.62918
```

```
print(mselas_tel)
```

```
## [1] 13.65177
```

```
#MSE comparison
```

```
library(nnet)
```

```
## Warning: package 'nnet' was built under R version 4.2.2
```

```
#18 MSE
for(i in 1:62){

  set.seed(4521)

  train_lin <- nnet(traino[,1]~., data=traino, size=i, linout=TRUE, trace=FALSE)
#calculating mse

  mse_nnet_lin <- mean((testo[,1]-(predict(train_lin, newdata=testo))))^2)
  print(paste("Number of hidden layer variables:", i))
  print(paste("MSE:",mse_nnet_lin))

}
```

```
## [1] "Number of hidden layer variables: 1"
## [1] "MSE: 83.6400769198489"
## [1] "Number of hidden layer variables: 2"
## [1] "MSE: 83.6400769165083"
## [1] "Number of hidden layer variables: 3"
## [1] "MSE: 47.977940044737"
## [1] "Number of hidden layer variables: 4"
## [1] "MSE: 39.1091722928104"
## [1] "Number of hidden layer variables: 5"
## [1] "MSE: 39.1949923948665"
## [1] "Number of hidden layer variables: 6"
## [1] "MSE: 36.4446841679003"
## [1] "Number of hidden layer variables: 7"
## [1] "MSE: 44.1099461071932"
## [1] "Number of hidden layer variables: 8"
## [1] "MSE: 33.8195759031686"
## [1] "Number of hidden layer variables: 9"
## [1] "MSE: 51.0366464010504"
## [1] "Number of hidden layer variables: 10"
## [1] "MSE: 37.9875291410755"
## [1] "Number of hidden layer variables: 11"
## [1] "MSE: 45.7785498364663"
## [1] "Number of hidden layer variables: 12"
```

```

## [1] "MSE: 38.178487621891"
## [1] "Number of hidden layer variables: 13"
## [1] "MSE: 28.7055049873686"
## [1] "Number of hidden layer variables: 14"
## [1] "MSE: 37.6641341614092"
## [1] "Number of hidden layer variables: 15"
## [1] "MSE: 40.6919007497503"
## [1] "Number of hidden layer variables: 16"
## [1] "MSE: 39.4461489344258"
## [1] "Number of hidden layer variables: 17"
## [1] "MSE: 34.2971207870739"
## [1] "Number of hidden layer variables: 18"
## [1] "MSE: 29.1344136885996"
## [1] "Number of hidden layer variables: 19"
## [1] "MSE: 34.8166165041897"
## [1] "Number of hidden layer variables: 20"
## [1] "MSE: 31.8381340782264"
## [1] "Number of hidden layer variables: 21"
## [1] "MSE: 36.5453890354278"
## [1] "Number of hidden layer variables: 22"
## [1] "MSE: 40.1162339270383"
## [1] "Number of hidden layer variables: 23"
## [1] "MSE: 29.1984491746797"
## [1] "Number of hidden layer variables: 24"
## [1] "MSE: 30.6209196491134"
## [1] "Number of hidden layer variables: 25"
## [1] "MSE: 39.0009897390151"
## [1] "Number of hidden layer variables: 26"
## [1] "MSE: 37.6301044930295"
## [1] "Number of hidden layer variables: 27"
## [1] "MSE: 27.4565021846831"
## [1] "Number of hidden layer variables: 28"
## [1] "MSE: 25.2400082698274"
## [1] "Number of hidden layer variables: 29"
## [1] "MSE: 34.7757955174109"
## [1] "Number of hidden layer variables: 30"
## [1] "MSE: 31.965064513475"
## [1] "Number of hidden layer variables: 31"
## [1] "MSE: 28.8816764649627"
## [1] "Number of hidden layer variables: 32"
## [1] "MSE: 34.0890677423365"
## [1] "Number of hidden layer variables: 33"
## [1] "MSE: 26.4959629224854"
## [1] "Number of hidden layer variables: 34"
## [1] "MSE: 28.4237369161042"
## [1] "Number of hidden layer variables: 35"
## [1] "MSE: 32.4164342387956"
## [1] "Number of hidden layer variables: 36"
## [1] "MSE: 20.1272287137891"
## [1] "Number of hidden layer variables: 37"
## [1] "MSE: 26.926865480393"
## [1] "Number of hidden layer variables: 38"
## [1] "MSE: 24.0283207377453"
## [1] "Number of hidden layer variables: 39"
## [1] "MSE: 19.03258233277"
## [1] "Number of hidden layer variables: 40"

```

```
## [1] "MSE: 31.2072052578396"
## [1] "Number of hidden layer variables: 41"
## [1] "MSE: 23.5092175915864"
## [1] "Number of hidden layer variables: 42"
## [1] "MSE: 24.1495351327313"
## [1] "Number of hidden layer variables: 43"
## [1] "MSE: 25.0944071760572"
## [1] "Number of hidden layer variables: 44"
## [1] "MSE: 20.5156309938443"
## [1] "Number of hidden layer variables: 45"
## [1] "MSE: 21.8931373664154"
## [1] "Number of hidden layer variables: 46"
## [1] "MSE: 25.0687371973734"
## [1] "Number of hidden layer variables: 47"
## [1] "MSE: 21.0506682210763"
## [1] "Number of hidden layer variables: 48"
## [1] "MSE: 27.4502377484659"
## [1] "Number of hidden layer variables: 49"
## [1] "MSE: 32.6795831213123"
## [1] "Number of hidden layer variables: 50"
## [1] "MSE: 25.5112208474508"
## [1] "Number of hidden layer variables: 51"
## [1] "MSE: 23.0842789858608"
## [1] "Number of hidden layer variables: 52"
## [1] "MSE: 22.5675262055132"
## [1] "Number of hidden layer variables: 53"
## [1] "MSE: 34.4376773022187"
## [1] "Number of hidden layer variables: 54"
## [1] "MSE: 25.7537544879999"
## [1] "Number of hidden layer variables: 55"
## [1] "MSE: 22.6121330749616"
## [1] "Number of hidden layer variables: 56"
## [1] "MSE: 34.4177773681845"
## [1] "Number of hidden layer variables: 57"
## [1] "MSE: 26.3048853995866"
## [1] "Number of hidden layer variables: 58"
## [1] "MSE: 22.8962963717843"
## [1] "Number of hidden layer variables: 59"
## [1] "MSE: 24.3579486898073"
## [1] "Number of hidden layer variables: 60"
## [1] "MSE: 29.387052337612"
## [1] "Number of hidden layer variables: 61"
## [1] "MSE: 28.3184113954799"
## [1] "Number of hidden layer variables: 62"
## [1] "MSE: 24.6092191010166"
```

##Pending 1. np specification test

```
# X <- data.frame(df$Adult.Mortality,df$infant.deaths,df$Hepatitis.B,df$BMI,df$under.five.deaths,df$Polio,
#
# result_npcms <- npcmsTest(model=model_lm, xdat=X, ydat=df$Life.expectancy) #33Hours to run

# result_npcms

# objects()
# find("result_npcms")
```

```

#
# saveRDS(result_npcms,"result_npcms.rds")
result_npcms <- readRDS("result_npcms.rds") #PreTrained Model
summary(result_npcms)

##
## Consistent Model Specification Test
## Parametric null model: lm(formula = Life.expectancy ~ Adult.Mortality +
##                               infant.deaths + Hepatitis.B + BMI + under.five.deaths
##                               + Polio + Diphtheria + HIV.AIDS + GDP +
##                               thinness..1.19.years +
##                               Income.composition.of.resources + Schooling, data =
##                               df, x = TRUE, y = TRUE)
## Number of regressors: 12
## IID Bootstrap (399 replications)
##
## Test Statistic 'Jn': 21.17521    P Value: < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Null of correct specification is rejected at the 0.1% level

```

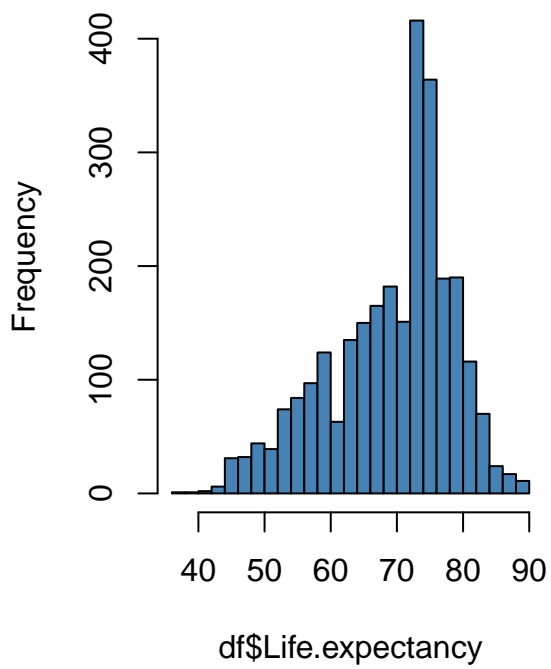
## 2. Visualizing bimodal distribution of yS

```

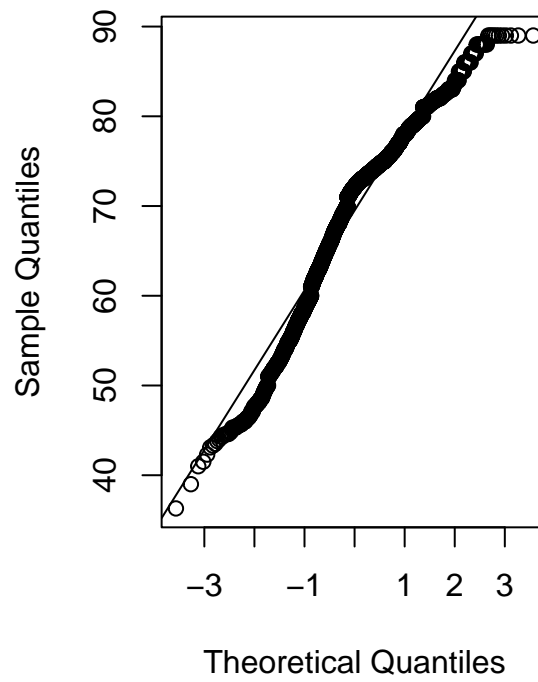
#Histogram & QQPlot
par(mfrow=c(1,2))
hist(df$Life.expectancy, col='steelblue', main='Life.expectancy_Histogram',breaks = 35)
#not really a good "bell-shape"
qqnorm(df$Life.expectancy, main='Life.expectancy_QQplot')
#S #most of the data is not fall along a straight diagonal line
qqline(df$Life.expectancy)

```

Life.expectancy\_Histogram



Life.expectancy\_QQplot



*#Both are indicating that our predict variable Y "df\$Life.expectancy" is not normally distributed*